Making electronic resource work for humanities scholarship

Hockey, Susan

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Making Electronic Resources Work for Humanities Scholarship

Susan Hockey (University of Alberta)[*]

Much humanities scholarship consists of the interpretation of source material. It is about illuminating and explaining those sources, and making connections or links between them. Traditionally these interpretations are published as monographs or journal articles. Other scholars may challenge these interpretations, publishing their own explanations with reference to earlier ones. Humanities scholarship is thus firmly grounded in critical thinking and assessment. In making electronic resources work for humanities scholarship, ways must be found to facilitate this kind of critical thinking and to represent multiple views of the same material.

Until recently much scholarship in the humanities has focussed on textual sources of many kinds including literature in the form of prose, Verse and drama; historical documents such as charters, newspapers, letters, diaries etc; spoken texts, that is transcriptions of conversations, speeches etc; dictionaries both historical and modern; and secondary material such as joumals and monographs. These texts are characterized by their complexity which must be represented and modelled effectively if the texts are to be used for high quality electronic scholarship. Electronic humanities texts must also be long-lasting and be capable of being studied by different scholars for different purposes in literary, historical, linguistic and cultural research.

Introducing technology to humanities scholarship poses interesting challenges. It can be time-consuming because the leaming curve is steep, but complex material requires complex software which must be mastered fully. Technology is subject to constant changes but must accommodate scholarly projects which in some cases last for many years. It is perhaps therefore inevitable that one common use of new technology is to deliver the old, that is to use the speed of the network to provide rapid access to print material. In this approach, which I hope is a transitory phase, the text can be described as

---

[*] Protokoll des 74. Kolloquiums über die Anwendung der EDV in den Geisteswissenschaften an der Universität Tübingen am 5. Dezember 1998.
Beginnend mit dem Kolloquium vom 7.2. 1998, stehen die Protokolle auch im INTERNET zur Verfügung und sind unter folgender Adresse erreichbar: http://www. uni-tuebingen.de/zdv/zrlinfo/kolloq.html.
Ähnlich wie es mit Beiträgen in den elektronischen Listen H-Soz-u-Kult und H-AHC geschieht, macht die allgemeine Zugänglichkeit der Protokolle im INTERNET es in Zukunft möglich, nur noch ausgewählte Beiträge aus den Protokollen in der Sektion »Humanities Computing« der Zeitschrift HSR abzudrucken.

»dead«, since it can only be searched by tools which have been designed for print media.

The model of print has developed over five hundred years. It is designed for the physical structure of the book with sequential pages supplemented by organizational aids such as the back of the book index, footnotes and bibliographies. Typography is a crucial feature of paper and books, and its function is to aid the reader by reinforcing what the text says. Typography is an important feature in current word processing programs which create electronic text for the purposes of producing print, even though typography is ambiguous for any other kind of computer processing. Within the same text italics can represent titles of books, foreign words or emphasized words. A reader has the intelligence to distinguish these. A computer program does not, unless it is supplied with more specifc information. However, most electronic text created today is via word processing programs which can only provide typographic encoding. This is likely to remain the Gase until there is a much more wide-spread appreciation of what can be done with electronic text and more widely available tools for creating other kinds of encoding.

Our understanding of electronic technology is nothing like as complete as our understanding of print technology, but we are becoming more aware of the opportunities afforded by this technology and what we might have to do to take advantage of these opportunities. Electronic information, or »live« text, is obviously much more flexible than print. It can be searched for any word or phrase, not just via a back of the book index. Individual pieces of electronic information such as paragraphs can be accessed separately. It is possible to make links between pieces of information, to merge pieces of information and to change information. Because of this flexibility and mutability of electronic information, the book-like form is no longer the only model. Other models may facilitate some kinds of humanities research, and their exploration is the subject of much research in humanities computing.

Electronic texts have traditionally been used in the humanities for manipulation and analysis, and for the production of printed scholarly editions. The advent of CD-Roms and now the World Wide Web have recently turned the focus more towards access and to preservation, to provide what can loosely be called »electronic publishing«. The requirements of these new »uses« of electronic texts as well as these of the earlier applications need to be considered in the design and development of electronic text markup systems and software. The Web provides a common, although rather deficient interface. The samecannot be said for CD-Roms where it is very easy for a new project to succumb to what has been called the »make a CD-Rom« syndrome where the project begins by choosing some proprietary software usually based on a higher priority for screen display than for functionality. The project then goes on enter data into the software's proprietary format only to find that the software developer has gone out of business before the project is finished.

Manipulation and analysis tools such as concordances and word counts have been used successfully for a variety of applications such as the study of style, themes, linguistic and philological features. At present, these applications do not appear to fit well into current trends in literary criticism, particularly in North America, but the humanities computing community with the expertise it has developed in handling complex electronic texts, has muck to offer the developers of electronic publishing and delivery systems which are very popular at present. Access is the focus of many current projects which aim to make available over the network material that was previously difficult to get to. Electronic access allows many people to work with the same object, and it also allows multiple routes into the same material. The question then arises of how best to facilitate that access. What descriptors are needed to help locate the electronic information and what functions are required for the delivery system? Other current projects focus on preservation and are attracted to digital media rather than microfilm because of the potential of electronic access, and also because electronic information does not degrade when it is copied. Many exact copies can be made and stored in different places. The very idea of digital preservation can be considered rather contradictory, because preservation implies that something is fixed, but by its very nature electronic information is not mutable and not fixed.

A summary of the current picture of electronic text technology in the humanities may Look something like this. Individual scholars working on their own research and teaching projects form one group of users. These are based in many different places and are most often working on a small number of texts in great detail, using traditional humanities computing text analysis methodologies. Library and publisher-based projects form another group of users. These are often larger projects which are putting many different texts into the same delivery system. The focus is less on fine detail and more on making larger amounts of text available for wider use over the network. It is also true that the Web is really becoming the operating system for most Computers and future development will be concentrated on the network and the Web.

The humanities scholar of the future will have access to digital objects of many different kinds on the network. It is not yet clear exactly what these objects will look like, but the need for a common interface and set of software tools is obvious. It is also clear that the future environment will be mixed. Print and other media will always be around and need to be accommodated in any future scenario for scholarship.

Digital imaging is also now making a significant contribution to humanities scholarship. It is one method of preservation which also enables remote access to material. The recent reduction in hardware costs and advances in compression techniques have now made large-scale high resolution imaging projects possible. Given the cost of taking the object to the camera or Scanner and setting it up, it makes sense to digitize at the highest resolution possible.

Image enhancement techniques make it possible to read previously illegible portions of manuscripts, and morphing techniques that can transform images have contributed to some projects. However images need text in order to be useful. Text in the form of descriptors or metadata is necessary to help locate the image. Annotations to images must be in the form of text and text is also needed to explain why an image is linked to something else. It follows that text associated with images should be subject to the same concems as electronic text and it is here where image projects can build on the expertise developed over many years of electronic text projects.

It took many years of work with electronic texts to recognize that a key issue for the longevity of data and projects is to keep the text separate from the Software. This implies ASCII files with markup that allows for many different purposes or applications for the text. Markup puts intelligence in the text and provides information to the Software by making explicit for Computer processing things which are implicit for the human reader. In pre-SGML days, two different types of markup were prevalent. Typographic markup schemes ranged from WYSIWYG word processors to embedded formats such as TEX, TROFF and proprietary schemes. Markup schemes for analysis programs were intended only for analysis. Many of these schemes used one syntax for references and another syntax for textual features making it difficult for an encoder to decide how to represent some kinds of features. None of these earlier schemes have adequate facilities for extension or for linking. Since it is much more powerful than these earlier schemes, SGML provides one syntax for representing all kinds of markup, thus making it possible to carry out many different processing functions on the same text. Because of its use of a document type definition (DTD), SGML also facilitates the processing of electronic texts. The DTD can be used for validating markup and it assists general purpose SGML software to carry out the same processing functions on different sets of markup tags.

The kind of descriptive markup provided by SGML is very important for scholarly applications, because it permits the definition of a model which corresponds to the text rather than forcing the text to fit an existing model. It thus avoids the loss of information likely to occur when data is entered into pre-defined and inflexible structures. The richness of SGML makes it possible to encode scholarly interpretation and to add new markup for different, and possibly opposing theoretical approaches to the same text. The re-usability of the text is another strong argument for SGML, since ultimately this leads to savings in money and time. The one disadvantage of SGML for scholarly applications is its inability to handle overlapping structures easily, other than with the Concur function which, as far as I am aware, has not been implemented in any widely-used SGML software. SGML assumes that the document is one single tree structure, whereas most, if not all, humanities texts can be viewed as multiple overlapping trees. Several projects, and notably the

Text Encoding Initiative, have developed ways of getting round the problem of overlapping structures, but most of these are somewhat clumsy to implement.

However, SGML is not particularly easy to use and it is not directly accessible via the World Wide Web. This has led to the development of the eXtensible Markup Language (XML) which is a cut-down version of SGML intended to work directly on the Web with the next generation of Web browsers. XML allows the user to define his or her own markup tags and to provide information to the Web browser on how to display these tags. Over the last year, a very broad base of interest in XML has grown very quickly in commercial and academic communities and this is being accompanied by the introduction of more software tools.

Ever since hypertext was first introduced more widely, humanities scholars have been interested in possibilities afforded by hypertext for linking pieces of information and modelling the kinds of connections which form the Basis of much humanities scholarship. A methodology is needed for expressing what is linked to what and, importantly for scholarship, why that link is being made. SGML and XML provide some excellent facilities for linking, since they essentially identify pieces of information marked by encoding tags. SGML and XML have mechanisms for encoding links between the pieces of information and can also encode ways of saying why those links have been made. Structures such as these provide what Yuri Rubinsky called the »underground tunnels« that make the pieces of a textbase work together.(1) They are the framework on which the scholarship can be built.

SGML can be used to encode anything, including material in other formats. It can provide the envelope of text that makes images, sound and other multimedia formats work effectively. It is thus not an alternative to Acrobat, PostScript and similar formats, but can function as a way of linking information in different formats. Apart from the Text Encoding Initiative, the Best-known SGML DTD for the humanities is the Encoded Archival Description (EAD: http://lcweb.loc.gov/ead/) which is used by Achval fnding aids. Finding aids are a good model for SGML because they are essentially hierarchic in nature with items in folders which are in boxes which are in series. It is possible to link EAD finding aids which point to the material to representations of the material itself encoded in the TEI or another SGML DTD.

Software for SGML has lagged behind the development of DTDs, but more tools are now emerging to cover the needs of document creation or conversion from legacy data, and the requirements for document browsing, retrieval or other manipulation. In my view, SGML/XML provide an excellent basis for electronic publishing since they provide the hooks needed to privilege certain navigation routes through the material. This is especially true for electronic scholarly editions where the editor must decide which routes are appropriate and point the user in the direction of these routes. The multiple routes made possible by SGML/XML can make possible new forms of scholarship which

are not yet fully understood. Pieces of information encoded in SGML/XML can be put together in many different ways. This means that the overall framework for a scholarly publications is no longer a single linear sequence of text with footnotes etc, but multiple pieces of information assembled and put together as the editor or user of the publication desires.

I am participating in two projects which are carrying out experiments in the delivery of pieces of scholarly information encoded in SGML. The Model Editions Partnership (MEP: http://mep.ela.sc.edu/ is developing a set of models for electronic documentary editions, material which forms the basic sources for American history. The goal of the MEP is to advance our understanding of what electronic documentary editions might look like and to create some samples of test material that Show different approaches to creating these editions. The MEP project includes seven documentary edition projects and it began by defining a prospectus for electronic documentary editions in collaboration with the partner projects. The prospectus outlines several basic principles, the first being that an electronic edition should to maintain current standards of scholarly editorial excellence. The other principles are that electronic editions should: facilitate changes in scholarly editorial practice; allow post-publication enhancements of editions; allow multiple forms of publication; and conform to relevant standards for electronic text, images, and other material.

The Orlando Project (http://www.ualberta.ca/ORLANDO) at the Universities of Alberta and Guelph is more ambitious. It is creating an integrated history of women's writing in the British Isles in electronic and printed form. Unlike most other humanities computing projects, Orlando is not encoding existing texts, but is writing new material (biographies of women writers, notes about other historical events, discussions of the writing history of each women author). All the material is encoded in rich SGML structures that incorporate detailed literary, historical and critical interpretation. This makes it possible to select from across the Orlando textbase pieces of SGML-encoded information that relate to women's education or to medical conditions or to travel or to political activities. All Orlando DTDs contain items for an overall chronology of women's writing in the British Isles. The SGML encoding makes it possible to select chronology items in many different ways, to sort these items and thus prepare selective chronologies for specific time periods, keywords or many other topics.

Depending on the nature of the work, startup costs for SGML can be higher than for some other methodologies. But SGML has many advantages to outweigh these costs. It forces members of a project team to think out clearly what they want to do while still leaving avenues open for extension and revision later on. More than anything it can be considered an investment for the future. SGML-encoded text will last for a long time and it can be enhanced as needed. Attention must now turn to making SGML work better. This means

research an how to facilitate inserting markup, how to add links (semi-)automatically and how to deal with multiple hierarchies. It also means research towards defining in more detail what kind of functions humanities scholars would like to carry out an SGML-encoded texts and building prototypes to test these functions. And, to ensure that the wheel is not re-invented, it means making known lessons learned from other projects, even if these projects have not turned out to be very successful.

A number of topics are crucial in planning for the future. The first of these, longevity of the data, has been discussed earlier in this paper. Data creation is expensive and so multi-purpose data that can migrate easily from one system to another makes sense from all respects. The second topic, metadata or data about the data, has been the subject of much research during the 1990's, once the need for it was clearly recognized. In the humanities, metadata is very important for a number of reasons. There is a need to know the source of an electronic text, the principles that governed the digitizing or encoding of the text, any revisions that have been made to the text and by whom they were made, and the size and scope of the text. This metadata should be in a form that is useful to both humans and computer programs, and its format should therefore be structured in some way. A means must also be found to work towards some common descriptor terms that can work across projects. The TEI provided one of the earliest metadata schemes for the humanities. Its electronic document header incorporates most, if not all, of the requirements listed above, and, because it is SGML, it can be processed by the same program that processes the rest of the text. The aims of the Dublin Core (http://purl.oclc.org/dc/) are somewhat different. It is a simple element set designed to aid resource discovery on the Web. It includes fifteen optional elements holding basic information which can be provided by the creators of digital objects. Another metadata structure is of course the library catalogue. This now has a field that links to a resource on the Internet, but it provides only a pointer to the resource, with very little information about the electronic properties of the resource.

There is a perception, at least, that software tools have not kept up with developments in markup technologies. Software with excellent functionality does exist – TuStep is a notable example – and a means must be found to make a very broad community aware of the potential afforded by good text processing tools for the humanities. Basic humanities text analysis tools of concordances and word counts also need to be enhanced by some of the tools being developed in computational linguistics so that more intelligent searching can be performed for concepts, not sequences of letters. More generally, I would like to see more research on the possibilities afforded by the delivery of SGML/XML-encoded text via the network so that the expertise developed in standalone text analysis systems can be incorporated into the electronic publication and delivery of scholarly humanities material. I think it is crucial

for potential users to be involved in the design of these systems. This will not only ensure that user needs are met, but it will also help to create a market for software systems for the humanities. Furthermore, experience has illustrated the difficulties and costs of support for complex software tools that are freely available. A way must be found to provide these tools at a modest cost, but one which will help to ensure the long-term viability and development of the software.

The Web is now very much the focus of computing, but CD-Roms are widely used for distribution of electronic information. With a wellthought out markup scheme it is possible to use both methods of distribution and not to view a CD-Rom as the ultimate product of a project. Since they are physical objects, CD-Roms fit better into some current procedures developed by libraries and by publishers for handling materials. However, updates are difficult to manage, since it is usually necessary to send a new CD-Rom to all purchasers. Libraries have also begun to find that support for a collection of CD-Roms can be expensive since the collection may include many different user interfaces and also technical incompatibilities. A single Web-based interface to collections of material is much easier to support and also works well for updates which can be controlled centrally. However, the current Web technology and HTML are weck for searching and, for anything other than very simple Web pages, require a clumsy interface to a more sophisticated back-end search engine. XML offers much more potential here but it is early yet to see how this will work out in practice. Nevertheless, the most appropriate way forward seems to be subscription-based services over the Web.

Questions also arise about whose role it is to develop electronic products for the humanities. The book publishing model is fairly simple. An academic author prepares a text over which he or she has intellectual control. The role of the publisher is to prepare the printed version, and distributeit by selling it to individuals and libraries through recognized routes such as bookshops and mailings. The publisher also serves as a gatekeeper, giving a seal of approval to the book and on acknowledgement of the value of the book's intellectual content. A number of book publishers have now started electronic publications in which their role is somewhat different. Some have initiated publications in-house and have developed these publications themselves either as extensions of their existing book publications (usually reference works) or as new publications. In either case they have more control over the intellectual content of the publication. They have also found that marketing electronic publications is rather different from marketing books and that costly demonstrations are often needed, rather than the mere mailing of leaflets. The cost of supporting users must also be built into any budget for electronic products and this usually means hiring or contracting technical staff. In contrast, Software publishers have more experience of the nature and problems associated with electronic information, but, in order to target the humanities scholarly community, their

marketing must also show a clear awareness of the scholarly content and an understanding of what humanities scholars do. Otherwise, they will face credibility problems.

Electronic Information is also creating a new role for academic libraries which are taking on the role of the middleman in the organization and dissemination of electronic information across campuses. Some libraries are building up collections of electronic texts and providing access to them via a unified Web-based interface to a search engine. This implies taking on some responsibility for the intellectual content of the material since intellectual access to the texts is controlled by the functions provided by the search engine and the user interface. Whoever builds the Index and designs the user interface must decide whether every word is to be indexed and how items that have been retrieved are presented to the user. Libraries are also beginning to digitize material from their own collections and to provide access to this via Web-based finding aids.

Humanities computing has made considerable progress during the fifty or so. However, much remains to be done to work towards a common methodology which allows for the many different and complex types of material that are the subject of research in the humanities and can also leave open the possibility of new and as yet undiscovered avenues of exploration. I would like to see more experimental projects like the Model Editions Partnership whose objectives are not just to make more Information available electronically but also to experiment with methodologies and to publish critical assessments of how well these methodologies have met the scholarly and technical requirements of the project. We need to build on the lessons learned by others and, in the current state of our knowledge, in many ways these lessons leamed in creating and using new resources are as important as the resources themselves.

Notes

1) Yuri Rubinsky, »Electronic Texts the Day After Tomorrow«, p. 5-13 in Visions and Opportunities in Electronic Publishing: Proceedings of the Second Symposium, December 5-8, 1992, edited by Ann Okerson, Association for Research Libraries, also available at http://arl.cni.org: 80/scomm/symp2/rubinsky.html.